Tutorial – 1
(Pairwise Sequence Comparisons)

9th April, 2021

Tutorial Using EMBOSS:

http://www.bioinformatics.nl/cgi-bin/emboss/

Since 2002, beta coronaviruses (CoV) have caused three zoonotic outbreaks, severe acute respiratory syndrome coronavirus, SARS-CoV, in 2002–2003, Middle East respiratory syndrome coronavirus, MERS-CoV, in 2012, and the newly emerged coronavirus, named SARS-CoV-2 in late 2019. CoV uses its spike glycoprotein (S), the main target for neutralization antibody, to bind its receptor, and mediate membrane fusion and virus entry. To understand the biology of SARS-CoV-2, let's find out its relatedness with the earlier CoV sequences and find the closest homologs.

1. **DotPlot Analysis** using (a) Dottup – k-tuple search, and (b) Dotmatcher – sliding window

Run these programs for different k-tuples (Dottup) and different window sizes and thresholds (Dotmatcher).

Given are the sequences of spike glycoprotein (both DNA and protein) of the following: (1) SARS-CoV (2003), MERS-COV (2012), and (3) SARS-CoV2 (2019). Submit the results of Dottup and Dotmatcher and answer the following Qs:

(i) Identify SARS-CoV2 is similar to which of the earlier two viruses? (ii) Is it easy to identify the similarity using DNA or protein sequences? Give reasons. (iii) Submit the graphs and give the k-tuple values used, and window size and threshold values used.

2. (a) **Pairwise Alignment:** Perform pairwise alignment of spike glycoprotein of SARS CoV2 with that of SARS-CoV, both at the DNA and protein level, using programs 'needle' and 'water'. Answer the following Qs.
   (i) What are percentage identity and percentage similarity at the DNA level and protein level? Which is larger and why give reasons.
   (ii) What is the difference between identity and similarity?
   (iii)Is there any difference in the global and local alignments of these two sequences?
   (iv)Submit the alignment giving the scoring scheme and gap penalties used.

   (b) **Pairwise Alignment:** Perform pairwise alignment of spike glycoprotein of

SARS CoV2 with that of MERS-COV virus, both at the DNA and protein level, using programs ‘needle’ and ‘water’. Answer the following Qs.

(i) Based on the sequence alignment, can you say that the two proteins are homologs, i.e., related?
(ii) Are you able to make this inference from the alignment of DNA sequences or protein sequences?

SARS-CoV-2 is a zoonotic virus that some jumped from animal species to humans.

3. Database search: Perform protein database search using spike glycoprotein of SARS CoV2 (Acc. ID: YP_009724390.1) as query and answer the following Qs:

(i) Which is the closest homolog of the query sequence?
(ii) Give the score, percentage identity, percentage similarity, length of the alignment, and the expect or e-value.
(iii) Do you find the spike glycoprotein of SARS-CoV as one of the hits? Do the percentage identity and percentage similarity results match with the alignment obtained using ‘water’? What is the significance of this alignment?
(iv) It was speculated that SARS-CoV2 has come from a bat. Do you find any relation of spike glycoprotein of SARS-CoV2 with that of bat SARS coronavirus spike glycoprotein? What is identity, similarity, length of the alignment, score, and e-value?

4. To identify from which species it might have jumped on to humans, perform BLAST database search using its genomic sequence (Acc. ID: NC_045512.2). [Hint: Set the number of target sequences to 1000 or more, and pay attention to query sequence coverage before reporting your results].
(a) Which BLAST nucleotide program would you use and why for genomic sequence comparisons?
(b) What do the top hits correspond to? Take any two top hits and analyze their alignment with the query and report the variation.
(c) List top 5 non-SARS-COV-2 sequence hits from different species and give the sequence coverage, %age identity, and e-value. Which coronavirus species is closest to SARS-COV-2? What do you infer from this result? Did SARS-COV-2 come from bat or pangolin, as earlier expected?

5. Find out the size of the protein database, UniProt, and nucleotide database, GenBank. Compute No. of matrix cells to be computed using DP for:
(i) Performing a search in the protein database, UniProt, and nucleotide database, GenBank and the time required assuming query sequence of length 1000 bases.
(ii) Comparing Human Chr 1 ~249Mbp with a query sequence of 1000 bases using DP, and comparing it with Chr 1 of Mouse (~195Mbp)? What is the memory or space requirement in the two cases?